Lexical Resources for Natural Language Processing Systems and the Problem of Polysemy

Cornelia Maria Verspoor
Microsoft Research Institute
Macquarie University
Sydney NSW 2109
kversp@cogsci.ed.ac.uk

Abstract

In this paper I evaluate extant lexical acquisition techniques for NLP systems with respect to the problem of polysemy. I argue that available lexical resources are inadequate for creating computational lexica capable of addressing the generativity of language use. I suggest several desiderata for lexical resources, to increase their utility for lexical acquisition in the context of polysemy: fully parsed corpora, and abstract semantic tagging of both verbs and nouns. These desiderata derive from a worked example of how lexical acquisition might proceed specifically in the context of handling the phenomenon of logical metonymy. I propose that both effective solutions to the word sense disambiguation problem and the development of adequate lexical resources for tackling this problem on a large scale must be informed by linguistic analysis.

1 Introduction

The computational lexicon is the fundamental repository of information about the primary component of language, i.e. words, and therefore critical for systems which aim to handle some aspect of natural language. Two key issues for the lexicon in natural language processing (NLP) tasks are lexical *representation* and lexical *acquisition*. In this paper I will consider the issue of acquisition in the light of the problem of polysemy. I start from the assumption that NLP systems will ultimately need to have the ability to identify the intended sense of a word at the same level of specificity as humans, and I will argue that (1) the way in which that problem is addressed must be informed by linguistic analysis and (2) the process of and resources for lexical acquisition must radically change in order to support effective word sense disambiguation. Polysemy, or word-level meaning ambiguity, imposes demands on the computational lexicon far greater than what can be extracted from existing lexical resources. I will end with a list of desiderata for future lexical resources which should be better structured for use in acquisition.

A sophisticated NLP system, i.e. one which aims to achieve deep interpretation of a wide range of texts and which can respond effectively to user input on the basis of that interpretation, will inevitably face the problem of word sense ambiguity as particular words can take on an almost indefinite number of subtle meaning variations. This ambiguity can derive from at least the following three sources: *homonymy* — a single lexical form might be associated with two entirely distinct meanings (e.g. *mogul*, an emperor, or *mogul*, a bump on a ski piste); *regular sense extension* — productive shifts in meaning which can occur with particular classes of words (e.g. a noun referring to an animal can be used to refer to the meat from that animal; *John walked his dog* and *We ate dog for dinner*); and *contextual modulation* — sense variation induced by a context, in terms of emphasising or de-emphasising various aspects of the sense (e.g. *door* as physical object in *John painted the door* or as aperture in *John walked through the door*). A distinction can be drawn between *established* senses of a word and *non-established* but licensed senses of a word. Machine readable dictionaries tend to focus on established senses, with no representation of productive, licensed meaning variations. Corpora are a potential resource for identifying a wide range of the senses which can be associated with a word, but there is no principled way of distinguishing established and non-established senses on the basis of a corpus and therefore no mechanism for identifying a productive process. This mechanism must derive from linguistic analysis.

While there has recently been growing interest in identifying and modeling the sources of this generativity within the lexical semantics community (Pustejovsky 1991, 1995; Copestake and Briscoe 1995;

inter alia) and there have been arguments made within the lexicographic community that the insights gained through this research should guide the design and development of machine-readable dictionaries (Atkins 1991; Atkins and Levin 1991), there has been little discussion of the incorporation of these insights into NLP systems. With this paper, I would like to open that discussion.

The paper will proceed as follows: in Section 2, I will introduce the phenomenon of logical metonymy and suggest what implications the polysemy implicit in this linguistic structure has for NLP systems; in Section 3, I will review current approaches to lexical acquisition, and highlight some of the problems with them; and in Section 4, I will discuss how lexical acquisition might occur within the framework of a particular analysis of polysemy and introduce the desiderata for lexical resources.

2 Logical Metonymy: A Question of Representation

Logical metonymy is the phenomenon that more meaning than is directly attributable to sentential components arises with certain verb/noun combinations. It is used to account for some verbs having alternate syntactic complement forms, but only a single semantic interpretation. Thus, the sentences in (1) express the same meaning, although in the (1b) no reading event is explicitly mentioned.

(1) a. John began reading/to read the book.
b. John began the book.

The systematic syntactic ambiguity of such verbs is handled via an operation of *type coercion* which shifts the object denotation of the complement NP (*the book*) to an event denotation, such that the logical forms for each verb-complement form will be identical.

The coercion which must occur to get the appropriate reading of (1b) requires that a missing element of meaning, i.e. the reading event, be introduced. Pustejovsky (1991) proposes that it is derived from a richly structured lexical semantic representation for nominals, the *qualia structure*, which captures defining attributes of the denotation of a noun. The attributes are represented by four roles: *constitutive*, specifying consituent parts of an object; *formal*, that which distinguishes the object within a larger domain; *telic*, the purpose or function of the object; *agentive*, how the object came into being. Type coercion looks to the qualia structure for something of the type required by the verb (an event in this case). For *book*, the reading event is available under its *telic* role and this provides the basis for coercion.

Verspoor (1997) argues on the basis of corpus evidence that logical metonymy is a phenomenon governed by conventionality. That is, there is an extremely narrow range of use of metonymic constructions in natural text, suggesting that type coercion is not a fully productive process. It is, rather, constrained by specification of usage conventions in the lexicon. In particular, the telic role of a noun may not in certain cases be lexically represented as it is not available to the process of coercion.

The implications of this type of polysemy for NLP systems are three-fold: first, meaning variations can derive from interactions among words and therefore cannot be fully specified in advance, second, rich lexical semantic structures are needed to account for the potential ambiguity of nouns like *book* and verbs like *begin*, and third, the precise instantiation of these semantic structures for a particular word depends on conventional knowledge of that word's usage and it is therefore not sufficient to simply identify a productive process without acknowledging the possibility of lexical exceptions. Each of these implications will be seen to pose a challenge to extant lexical acquisition procedures.

3 Lexical Acquistion: How It's Done

Many of the early NLP systems relied on hand-coding of the lexicon, but this was quickly realised to be problematic for the development of large-scale systems. Research turned to development of automated techniques for encoding of the lexicon. The initial attempts in this area were made by basing lexica on electronic versions of dictionaries, Machine Readable Dictionaries. However, more recently corpora have also been utilised as a source of information about frequency and co-occurrence information. I introduce and evaluate the achievements in what follows.

3.1 Machine Readable Dictionaries (MRDs)

Much effort has been invested in attempting to automatically extract information from MRDs which is then converted or incorporated into lexical entries for an NLP system. Initial attempts in this area focused on identification of morphological and syntactic information, but more recently there has been research into extraction of lexical semantic information as well. This research has largely concentrated on identifying taxonomies (e.g. Chodorow *et al.* 1985), although some work in identifying related sets of words has also been undertaken (e.g. Byrd *et al.* 1987).

However, several problems with using MRDs are repeatedly identified in the literature. They can be summarised as follows:

- Dictionaries are written with a human user in mind and using natural language; the definitions therefore make full use of the subtleties and ambiguities in natural language.
- Word meaning is divided into discrete senses which are treated as independent; relationships between words and word senses (including hierarchical relationships between senses) are only implicitly represented in the text of the dictionaries.
- Dictionaries are by their nature finite and can therefore not account for the generative nature of word use. Word senses are identified in advance, independent of specific contexts.
- MRDs often do not use consistent formats in their definitions; there may be typographical errors; definitions may be circular or have internal inconsistencies.

The amount of semantic information useful for NLP which has been automatically extracted from dictionary definitions is severely limited. MRDs do not explicitly represent lexical relations, and gaps easily filled in by a human user cannot be automatically identified or resolved. The construction of semantic taxonomies is clearly very important for NLP systems for capturing generalisations over semantic classes of words, but if these taxonomies are derived from a source which makes artificial divisions between word senses in some cases and conflates word senses which might have linguistically significant differences in other cases (such as for the distinct syntactic frames of *begin*), their utility for precise interpretation seems questionable.

MRDs are specifically deficient for addressing polysemy, as follows from the problems summarised above. They are fixed and finite, and words are listed in isolation of one another, so the interactions inherent in logical metonymy and contextual modulation cannot be identified and the full range of meaning variation is not reflected. All of the core information relevant for qualia structure may not be captured in a dictionary definition — although it is likely that the formal and constitutive roles may be found in the entry, it is less likely that agentive and telic information will be mentioned — and if it is present, it will not be present in a consistent format useful for automatic extraction. Conventional knowledge of word usage is not even implicitly represented, as MRDs are a secondary source of word knowledge and concentrate on conveying established senses. For these reasons MRDs are not adequate as a source of lexical knowledge for computational systems.

3.2 Corpus-based Acquisition

In contrast to dictionaries, text corpora are primary sources of information about language use. They support detailed studies of how particular words are used by providing extensive examples of natural language sentences in context (Atkins 1991). They can be analyzed using statistical techniques, to derive information about word frequency, to establish co-occurrence frequencies of pairs of words, and to identify sets of words with similar distributions (e.g. Church *et al.* 1994). Some research, e.g. Fukumoto and Tsujii (1995) and Zhai (1997), focuses on identifying semantic classes of words and homonymous words based on statistical clustering, following from the premise that semantically similar words appear in similar contexts. This research is to a certain degree promising for lexical acquisition, in that it establishes relationships between words as actually used *in context* in natural language texts. Furthermore, collocations and non-lexical units (e.g. idioms), a very important component of linguistic knowledge,

can be identified using statistical analysis. However, semantic distinctions can only be made at a coarse level of granularity due to the need for statistically significant differences; closely related uses of a single word and subtle meaning distinctions could not be automatically discovered.

In addition, corpora suffer from the fact that they consist only of surface words and no structure — in many corpora not even syntactic annotations are provided, much less clues about the meaning. The linguistic information necessary even for accurate classification of words into taxonomies is simply not available from a corpus. As concerns polysemy, the kind of information needed to build rich lexical semantic structures may be implicitly available in corpora because they provide a large body of evidence about word usage, but its extraction must be guided by specific linguistic hypotheses, as we will see below. Linguistic pre-processing will be necessary to convert corpora into a form useful for lexical acquisition, by identifying parts of speech, syntactic structures, morphologically related words, etc.

4 Lexical Acquisition: How It Should Be Done

The fundamental problem for extant automatic lexical acquisition techniques is that they assume clear divisions between word senses. However, word sense distinctions are not easily justifiable in isolation of particular contexts. These approaches seemingly deny the creative aspect of language use from the outset and will therefore always fall short of the ultimate goal of identifying the underlying principles of generative language use. NLP systems which require sophisticated language processing demand a framework which will accommodate the flexibility of language use and which will result in fine-grained interpretation. This framework can only come from linguistic theory. This theory can identify regularities in sense extensions, capture syntactic and semantic generalisations associated with particular groups of words, and guide lexical acquisition.

In line with the Atkins (1991) and Atkins and Levin (1991) arguments for development of MRDs, I would like to suggest that an adequate computational lexicon for NLP can only be established on the basis of top-down design derived from a linguistic theory in combination with bottom-up information derived from corpora about specific usage of language. The linguistic theory should guide the search for particular kinds of information in corpora, and should establish criteria for structuring and interpreting the data. The information derived from corpora might include sense frequency information, co-occurrence relations and collocations. It should also include idioms and representation of proper nouns, which establish contexts in which a word can take on non-compositional meanings.

Let us consider, as an example of how linguistic theory could guide lexical acquisition, how the automatic acquisition of the knowledge relevant to logical metonymy might proceed, given the theoretical analysis presented in Section 2. This involves acquiring the telic and agentive components of the qualia structure representation for nouns, and encoding the associated conventionality. A certain amount of the work of acquiring qualia structure will be shown to be feasible via automatic means, given a particular corpus structure, yet some of it still must be built up by hand, as will be pointed out in step (2c).

1. The values of potential agentive and telic roles must be identified for every artifact-referring noun. This involves (by hypothesis) identifying the verbal relations in which the noun most frequently plays a role. Two kinds verbal relations are most of interest:
   (a) The agentive role of a noun is likely to be the most frequent occurrence of a *creation* verb in which the noun is the created entity. For example, *bake* would be assumed to be the value of the agentive role for *cake* if *bake cakes* appears more frequently in the corpus than any other creation activity involving *cake*.
   (b) The telic role of a noun is likely to be the most frequent occurrence of any *non-creation* verb in which the noun is acted upon in some way. For example, *read* would be assumed to be the value of the telic role for *book* if *read books* appears more frequently in the corpus than any other non-creation activity involving *book*.

2. Instances of logical metonymies must be identified and analysed.
   (a) Pick out instances of an aspectual verb followed by a complement noun (phrase) which does not refer to an event. The restriction to complements means that this step can only occur after deep parsing has established the structure of sentences in the corpus.
   (b) For those nouns which don't participate in logical metonymies in the corpus, propose that their telic role is not accessible to the process of logical metonymy, and that therefore their telic role should not be lexically represented.
   (c) For those nouns which participate in logical metonymies in the corpus, attempt to identify whether the logical metonymies are agentive role-centred or telic role-centred, i.e. whether the ellided event is a creation or a non-creation event. This portion of the analysis involves extensive context-dependent interpretation and therefore does not seem to be possible automatically (it would involve the full power of a natural language understanding system). However, the preceding stages will have identified the relevant set of examples (likely to be small due to the restrictiveness of the phenomenon) in the corpus. As soon as a single non-creation metonymy involving a certain noun is found, it should be assumed that the telic role for that noun is represented and the next noun can be considered.
3. Add the potential agentive role to the lexical entry for each noun; add the potential telic role to the lexical entry only if there was evidence to do so found in step (2c).

The process described above requires a particular framework to be in place before such specific corpus analysis can proceed:

- An ontology must have been established which divides nouns and verbs into very general types, such as nouns which refer to artifacts, natural entities, events, and verbs which refer to *creation*/*non-creation* relations. During pre-processing of the corpus through parsing, these types should be clearly identified.
- General noun types must be defined for the relevant qualia structure roles. The agentive role, for example, is relevant to the nouns which refer to artifacts, while it is not for nouns which refer to natural entities. This will constrain the lexical structures to be created.
- A rigorous definition of each of the potential qualia structure roles must be presented in order to guide the search for their values for specific nouns.

The previous discussion gives some indication of what a corpus needs to look like in order to support the desired processing. Specifically, I suggest the following desiderata for a corpus useful for identifying semantic relationships:

- The corpus must have been parsed, in order to identify the structural relations between elements in sentences in the corpus. This process can determine, for example, when a verb is being used in the passive form, and which elements of a sentence correspond to syntactic and semantic arguments of the main verb.
- Verbs should be tagged according to their semantic type (which can depend on the results of the parsing). This information will be derived in a pre-existing ontology for verbs and the associated syntagmatic information represented there.
- Nouns in the corpus must be tagged according to their general semantic type (artifact-referring, event-referring, etc.). Again, this information will depend on a pre-existing ontology for nouns.

5 Conclusion

The development of a rigorous theoretically-derived analysis of polysemy is critical, due to the generative nature of word sense variation. No fixed set of word senses, such as those assumed by extant lexical

resources, can ever hope to be sufficient for addressing this generativity. However, in combination with a flexible theoretical structure, these lexical resources can be improved and together will ultimately allow computational systems to achieve the capability of handling the challenge posed by polysemy and the creativity of language use.

In this paper, I have examined lexical acquisition as is possible based on the current state of lexical resources. I argued that neither MRDs nor corpora can adequately support the creation of a generative lexicon. The extraction of information useful to advanced NLP tasks from lexical resources demands a certain level of linguistic sophistication both from the resources and from the framework of lexical structure which guides that extraction. Through the example of logical metonymy, we saw that the linguistic analysis of a phenomenon can guide lexical acquisition and will influence the structure demanded of lexical resources. Further investigations into the linguistic nature of polysemy will certainly lead to a desire for increased structure in lexical resources; finding ways to make lexical resources more informative will ultimately lead to more flexible NLP systems.

References

Atkins, B. (1991). Building a lexicon: The contribution of lexicography. *International Journal of Lexicography 4*(3), 167–191.

Atkins, B. T. S. and B. Levin (1991). Admitting impediments. In U. Zernik (Ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Chapter 10, pp. 233–262. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. Publishers.

Byrd, R. J., N. Calzolari, M. S. Chodorow, J. L. Klavans, M. S. Neff, and O. A. Rizk (1987). Tools and methods for computational linguistics. *Computational Linguistics 13*(3-4), 219–240.

Chodorow, M., R. Byrd, and G. Heidorn (1985). Extracting semantic hierarchies from a large online dictionary. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, University of Chicago, pp. 299–304.

Church, K., W. Gale, P. Hanks, D. Hindle, and R. Moon (1994). Lexical substitutability. In B. Atkins and A. Zampoli (Eds.), *Computational Approaches to the Lexicon*, Chapter 6, pp. 153–177. Oxford University Press.

Copestake, A. and T. Briscoe (1995). Semi-productive polysemy and sense extension. *Journal of Semantics 12*(1), 15–68.

Fukumoto, F. and J. Tsujii (1995). Representation and acquisition of verbal polysemy. In J. Klavans (Ed.), *Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*, Menlo Park, California, pp. 39–44. 1995 AAAI Symposium: AAAI Press. Technical Report SS-95-01.

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics 17*(4), 409–441.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Verspoor, C. M. (1997). Conventionality-governed logical metonymy. In H. Bunt, L. Kievit, R. Muskens, and M. Verlinden (Eds.), *Proceedings of the Second International Workshop on Computational Semantics (IWCS-II)*, Tilburg, NL, pp. 300–312.

Zhai, C. (1997). Exploiting context to identify lexical atoms – A statistical view of linguistic context. In *Proceedings of the International and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-97)*, Rio de Janeiro, pp. 119–129.